Similarity Avoidance in Bengali Fixed-Segment Reduplication

Sameer ud Dowla Khan
UC Los Angeles
sameer@humnet.ucla.edu
The question

- Similarity has been shown to be phonologically important in numerous recent studies.

- The phenomenon of similarity avoidance is seen in East Bengali fixed-segment echo reduplication.
  - Data on this phenomenon was gathered in an experiment.

- But how is similarity calculated?
  - Four theories of similarity were tested against the experimental results and compared to one another.
The alternation: East Bengali fixed-segment echo reduplication

- Fixed-segment reduplication is the copying all base material into the reduplicant, except for the fixed segment (McCarthy & Prince 1986).

- “Echo reduplication” is one instantiation of fixed-segment reduplication.

- East Bengali echo reduplication:
  - The reduplicant-initial segment is usually replaced with the default fixed segment /t/.
  - Alternate fixed segments, such as /f/, /m/, /z/, /p/, and /b/ are also attested.

- Using default fixed segment /t/:
  - pani 'water'
  - pani tani 'water etc.'
  - kaši 'cough'
  - kaši taši 'cough etc.'
  - loha 'iron'
  - loha toha 'iron etc.'
  - muri 'puffed rice'
  - muri turi 'puffed rice etc.'

- Using other fixed segments:
  - tægra 'cross-eyed'
  - tægra mægra 'cross-eyed etc.'
The phenomena: Identity- and similarity avoidance

- The choice of fixed segment in Bengali is subject to two types of cooccurrence restrictions:

  **Identity Avoidance**
  Reduplicants with a fixed segment identical to the segment it is replacing are rejected.

  **Similarity Avoidance**
  Speakers also reject reduplicants with a fixed segment, e.g. /t/, that is similar to the segment it is replacing e.g. /tʰ/.

- a. tajnna 'having pulled'
  *tajnna tajnna 'h. pulled etc.'
  tajnna majnna 'h. pulled etc.'

- b. majra 'having beaten'
  majra tajra 'h. beaten etc.'
  *majra majra 'h. beaten etc.'

- c. tʰajšša 'having stuffed'
  *tʰajšša tajšša 'h. stuffed etc.'
  tʰajšša majšša 'h. stuffed etc.'
Questions and preview

- **What is the basis on which speakers judge “similarity”?**
  - Does the lexicon play a role?
  - Do features and natural classes play a role?
  - Does the phoneme inventory play a role?

- **To follow:**
  - An experiment gathering data on similarity avoidance using native speaker judgments, and
  - Evaluation of four theories of similarity.
Experiment

- **Stimuli:** 60 native Bengali disyllabic roots.
  - **Identity condition:** 8 stimuli beginning with /t/.
  - **Similarity condition:** 29 stimuli beginning with consonants potentially considered similar to /t/ (e.g. aspirated /tʰ/, /d/, dental /t/, etc.).
  - **Control condition:** 23 stimuli beginning with other consonants.
- Produced by adult female native speaker in a sound-proof booth.
- Stimuli and multiple choices randomized for each speaker.
- No word included consonants similar to /t/ in non-initial position.
- Subjects: 30 adult native speakers of Bengali.
- **Presentation of stimuli:**
  - Word played aloud to subject.
  - Subject asked to repeat the word aloud with its reduplicant.
Experimental results

- The overall pattern followed the principles of identity- and similarity avoidance.
  - Bases with initial consonants such as /t/, /th/, /d/, and /tʃ/ took few reduplicants with fixed segment /t/.
  - Bases with initial consonants such as /l/, /m/, /p/, and /bʰ/ most often took reduplicants with fixed segment /t/.
  - Bases with other initial consonants show more variable behavior.
- The percentage of fixed segment /t/-use in echo reduplicants is inversely related to the similarity between /t/ and the base-initial consonant (see following graph).
Graph of experimental results

Fixed segment /t/-use in reduplicants

% of Reduplicants with FS /t/

More /t/-use in reduplicants

Presumed similar to /t/  Base-initial consonant  Presumed not so similar to /t/

/h/ is an outlier

Graph of experimental results

Fixed segment /t/-use in reduplicants

% of Reduplicants with FS /t/

More /t/-use in reduplicants

Presumed similar to /t/  Base-initial consonant  Presumed not so similar to /t/

/h/ is an outlier
Question and preview

- What theory could explain this data?

- Four theories are considered:
  - Lexical Cooccurrence Restrictions
  - Shared Natural Classes Metric
  - Relativized OCP Constraints
  - Feature Weighting
Theory I: Lexical cooccurrence restrictions (OCP)

- **OCP:** In the lexicon, identical (and similar) consonants tend to cooccur less frequently than more dissimilar consonants (McCarthy 1986)
- If these OCP restrictions in the lexicon are the only constraints productively applied in the grammar, speakers will judge similarity based on their lexicon
- **Prediction:** Bases with initial consonants that cooccur less often with /t/ in the lexicon will allow fewer /t/-reduplicants than other bases
- **Conclusion, if borne out:** Similarity is an entirely language-specific measure, based on the lexicon
Theory I: Implementation

- Using phoneme frequency and distribution data from Mallik et al. (1998), consonant cooccurrence with /t/ in Bengali tVCV and CVtV roots was calculated by means of observed/expected (O/E):

\[
\text{Observed \{t, C\} cooccurrence in roots} = \frac{\text{Observed /t/ occurrence in roots } \times \text{Observed /C/ occurrence in roots}}{\text{Total roots} \times \text{Total roots}}
\]

- O/E values less than 1 suggest the segments are subject to a cooccurrence restriction (OCP), and are thus being judged as more similar.
- Higher O/E values suggest the cooccurrence of the segments is favored in the language due to their dissimilarity (no OCP restriction).
No correlation between the predictions and the experimental data [$r^2 = .004$, $p > 0.81$]

This strongly suggests that the OCP effects in the Bengali lexicon are **totally unrelated** to the OCP effects in fixed-segment echo reduplication.
Theory II: Shared natural classes metric

- Frisch (1996) proposes that lexical and productive measurements of similarity are made by counting natural classes (see next slide for formula).
- This similarity measure has been shown in Frisch, et al. (2004) to be applicable to both the lexicon and grammar of Arabic.
- **Prediction:** Bases with initial consonants that share more natural classes with /t/ will take fewer /t/-reduplicants than other bases.
- **Conclusion, if borne out:** Similarity measurement has a universal component (features) and a language-specific component (the natural class inventory).
Theory II: Frisch et al.’s formula

- The similarity of a consonant C and /t/ can be measured using the following equation:

\[
\text{Similarity of } /t/ \text{ and } C = \frac{\text{Natural classes shared by } /t/ \text{ and } C}{\text{All natural classes relevant to } /t/ \text{ and } C}
\]

- Higher scores (approaching 1) represent more similar consonants, while lower scores (approaching 0) represent more dissimilar consonants.

- Bases that begin with consonants deemed more similar to /t/ should accept fewer /t/-reduplicants.
Theory II: Comparison with results

The shared natural classes metric is better at predicting most of the experimental results \( r^2 = .584, \ p < 0.01 \).

However, some important contrasts between coronal obstruents (e.g. /\text{th}/ vs. /d/, /\text{th}/ and /\text{t}/, etc.) are not predicted.
Theory III: Relativized OCP constraints

- Coetzee and Pater (2005) proposes an OT account of Muna:
  - OCP constraints against certain larger feature combinations
  - OCP constraints against smaller combinations thereof
  - General OCP constraint

- In the Muna analysis, this ranking derives from the lexicon by an algorithm based on the type frequency of lexical exceptions to weak OCP constraints.

- **Prediction:** Bases with initial consonants that share more combinations of features with /t/ will take fewer /t/-reduplicants than other bases.

- **Conclusion, if borne out:** *Similarity measurement has a universal component (features) and a language-specific component (lexicon).*
Theory III: Implementation

Since the Bengali lexicon is not correlated with the reduplication facts, the following OCP hierarchy is an attempt to model the similarity avoidance data with no reference to the lexicon:\(^1\)

\[
\]

\(^1\) This is the closest match between the relativized OCP theory and the observed /t/-use patterns. Other hierarchies considered yielded worse fits to the data.
Theory III: Comparison with results

- The hierarchy of OCP constraints posited here is a relatively close match to the experimental results \([r^2 = .717, p < 0.01]\).
Similarity of phonemes may be measured by counting up the features they share.

However, some features may be more considered important in particular languages.

Thus, when calculating the similarity of two consonants, certain features will be more heavily weighted than others.

**Prediction:** Bases with initial consonants that share a greater similarity weight with /t/ will take fewer /t/-reduplicants than other bases.

**Conclusion, if borne out:** Similarity measurement has a universal component (features) while allowing for language-specific weights.
Theory IV: Application

- Feature weights were calculated in the software program R (R Development Core Team 2005) by applying the following equation to fit to the observed reduplication data:

\[
P = \left(\frac{m!}{n!(m-n)!}\right) (1 - \text{sim}(C, t))^n \text{sim}(C, t)^{m-n}
\]

- Where \( P \) is the probability that base-initial \( C \) will cooccur with default fixed segment /t/ \( n \) times out of \( m \) trials, and \( \text{sim}(C, t) \) was calculated as:

\[
\text{sim}(C, t) = \exp\left(-\sum_{i=1}^{\text{# features}} w_i (1 - \delta_i(C, t))\right)
\]

\[
\text{Where } \delta_i(C, t) = 1 \text{ if } C \text{ and } /t/ \text{ agree on feature } i \text{ and } \delta_i(C, t) = 0 \text{ otherwise.}
\]

- Most features received the default feature weight (\( w \)) of 0.100
- The following four features were found to have heavier weights (\( w \)):

\[
\begin{align*}
\text{[voice]} & \quad 0.554 \\
\text{[distributed]} & \quad 0.400 \\
\text{[strident]} & \quad 0.249 \\
\text{[spread glottis]} & \quad 0.198
\end{align*}
\]

These four features turn out to be independently important in the language (see slide 24)
Theory IV: Comparison with results

- Feature weighting achieved the best match with the observed data \([r^2 = .855, p < 0.01]\).
Summary of modeling results

- The lack of any correlation between lexical coocurrence restrictions and /t/-use in reduplicants \( r^2 = .004 \) confirms that speakers do not judge similarity solely based on patterns in their lexicon.

- The correlation between shared natural classes and /t/-use in reduplicants \( r^2 = .584 \) cannot describe the particular behavior of the coronal consonants, suggesting that speakers do not judge similarity based on the number of shared natural classes.

- While the data can be closely predicted by positing relativized OCP constraints \( r^2 = .717 \), this requires the use of eight constraints that have no basis in the lexicon, and no predictions about similarity phenomena in other languages. It is unclear how speakers could acquire this grammar from independent sources.

- The theory that best fits the experimental data \( r^2 = .855 \) is one in which:

  \[ \text{Similarity is judged based on universal features assigned different weights.} \]
Further questions and a hypothesis

- If similarity is judged by assigning different weights to different features, where do these weights come from?
- Are the weights universal or language-specific?
  - If they are universal, then we predict that all languages would weight [voice], [distributed], [strident], and [spread glottis] above other features.
  - If they are language-specific, feature weights might be reflecting the relative importance of the feature in maintaining phonemic contrasts in the language.

Prediction: the better a feature is at maintaining phonemic contrasts in a given language, the heavier its weight.
A possible hypothesis (continued)

- The features that were weighted more heavily than others in Bengali include [voice], [distributed], [strident], and [spread glottis].
  - These very features effectively distinguish all 15 coronal obstruents in East Bengali, thus presumably carrying a large functional load in the language.

- If this data is representative of a larger pattern, we can predict that while phonetic features are universal, they have language-specific weights corresponding to the capacity of each feature to make phonemic contrasts.

- Data from productive similarity avoidance alternations in a variety of languages will be needed to test this.
References and acknowledgments

- **Frisch, Stefan, Janet Pierrehumbert, & Michael Broe (2004).** Similarity Avoidance and the OCP. *NLLT* 22.

I would especially like to thank my advisors, Kie Ross Zuraw, Colin Wilson (also my statistics consultant), and Bruce Hayes; my native speaker consultant, Farida Amin Khan; the UCLA Phonology Seminar; and the 30 subjects of my study.